Plausibility of Pathogenicity Based on Frequency (PloPaBoFy)

Context

In the process of undertaking a genetic mutation screen by gene sequencing (especially multi-gene screens) it is common to detect a variant that is of uncertain significance in relation to the disease or phenotype in question. Very often, a literature review reveals no information specific to the variant, and the only relevant information that can be found is that it has been detected in a certain number of individuals in a population-based cohort (e.g. Exome Aggregation Consortium [ExAC] project). This raises the following pertinent question: *Is it plausible, based on the cohort frequency alone, that a variant of this frequency could account for a proportion of individuals with the disease / phenotype in question*? Clearly – unless the disease / phenotype is typically caused by variants with greatly reduced penetrance – if the cohort frequency is higher than the disease frequency (i.e. prevalence), the answer to this question is: *No*! But what if the variant genotype frequency is significantly below the disease prevalence?

An ostensibly sensible rule-of-thumb might be to consider the specific undisputed pathogenic mutation that is most frequently detected in individuals with the disease in question. This approach assumes that if a newly detected variant is more frequent in the population-based cohort than that mutation, then it is unlikely to be of relevance. There are a number of problems with this assumption. Firstly, the accuracy of this assumption is dependent to some extent on the penetrance of that most common mutation. If that mutation is highly penetrant, it may be conceivable that a mutation that causes fewer cases of the disease / phenotype but with lower penetrance might have a higher population genotype frequency than the most common causative mutation. Secondly, the number of individuals in whom the most common pathogenic mutation is detected in a population-based cohort only represents one of many possible outcomes. A higher (or, indeed, lower) number may actually be a more likely outcome.

What is PloPaBoFy?

PloPaBoFy is a tool that is designed to help scientists interpret population-based cohort allele or genotype frequencies using a rational approach that avoids the shortcomings of the above-outlined, rather crude, approach. PloPaBoFy is a calculator that utilises the principles of Hardy-Weinberg equilibrium and the binomial distribution to determine the plausibility that a variant detected in a specific number of individuals in a population-based cohort could account for a specific proportion of variants or mutations that are causative for a specific disease or phenotype. Its usefulness is dependent on the scientist making reasonable (or, at least, conservative) assumptions about the disease or phenotype in question (for example, prevalence, penetrance, and genetic heterogeneity).

To use PloPaBoFy, users are required to input certain frequency information that relates to the genetic architecture of the disease or phenotype in question. This information includes the prevalence of the disease / phenotype, its inheritance pattern, and the expected penetrance of a causative genotype (whether hemizygous, heterozygous or homozygous). PloPaBoFy then utilises the Hardy-Weinberg principle to calculate the expected population allele frequency for a variant with the inputted profile. Users can also specify a certain percentage of the total allele frequency that PloPaBoFy considers. For example, it may be known or assumed that the maximum proportion of affected individuals accounted for by a single specific mutation (i.e. the most frequent disease-causing mutation) is 5%.

Once the expected population allele frequency has been calculated, the Hardy-Weinberg principle is again used to calculate the expected population frequency of all possible genotypes (e.g. homozygous wildtype [p^2], heterozygous [$2pq$], and homozygous mutant [q^2]). These genotype

frequencies are then assumed to represent the probability that a randomly selected individual will have either one of the three genotypes. A binomial distribution probability function is then used to calculate the distribution of the probability that a variant with the calculated genotype frequency profile will be detectable (either hemizygously, heterozygously, or homozygously) in (all possible values of) *x* of (a specified total of) *y* individuals in a population-based cohort. Once calculated, the user is able to input a cohort genotype count (i.e. i.e. a value of *x*) and PloPaBofy will return the probability that this number (or up to / at least this number) of individuals would be detected in a population-based cohort for a variant with the inputted genetic architecture profile (allele counts can also be inputted and an equivalent probability will be outputted). Accompanying the returned probability value will be an indication of the plausibility that a variant with the inputted cohort frequency could represent a variant with the inputted genetic architecture profile.

PloPaBoFy version 2 also calculates plausible ranges for the various parameters of the genetic architecture of the disease / phenotype. Modifications to any of these parameters in the genetic architecture input to a value within the indicated range will make the observed population cohort count plausible.

How to Use PloPaBoFy

PloPaBoFy can be accessed at http://www.genecalcs.weebly.com/plopabofy.html. There are two versions: The non-Express version is recommended for users that are unfamiliar with the program, as this version includes step-by-step instructions and full interpretation of the results. This version is also recommended where the genetic architecture of the disease / phenotype is less certain, as it allows users to input a range of values. The Express version performs exactly the same calculations. It is quicker to use, but does not include instructions or interpretation.

All yellow boxes require input from the user.

The PloPaBoFy inputs are divided into three sections:

1. ***General Information / Options*:** In the non-Express version, users are required to input an abbreviation for the disease / phenotype and input the gene symbol. This is not required in the Express version. In both versions, there is a grey table with yellow input cells that allows users to alter the *P* value thresholds for the plausibility conclusions. In the non-Express version, this table as at the top of the interface; in the Express version it is at the bottom of the interface. In most instances, the default values are recommended (i.e. 0 – 0.001 = 'highly implausible'; 0.001 – 0.01 = 'implausible'; 0.01 – 0.05 = 'possible (unlikely)'; 0.05 – 0.2 = 'plausible'; and 0.2 – 1 = 'highly plausible').
2. ***Genetic Architecture*:** This section allows users to input details about the genetic architecture of the disease or phenotype in question. These details include the inheritance pattern, prevalence, the expected penetrance of causative variants, and an estimation of the maximum proportion of cases or mutations that could be accounted for by any one variant. The values specified for these details are used to calculate an expected population allele frequency for a causative variant.
3. ***Observed Population Allele / Genotype Count*:** This section requires users to input an allele or genotype count for a variant in a population-based cohort. The size of the cohort must also be inputted. The number of females in the cohort is required only if the inheritance is X-linked and genotype input is selected.

The outputs of PloPaBoFy can be used to answer two opposite questions:

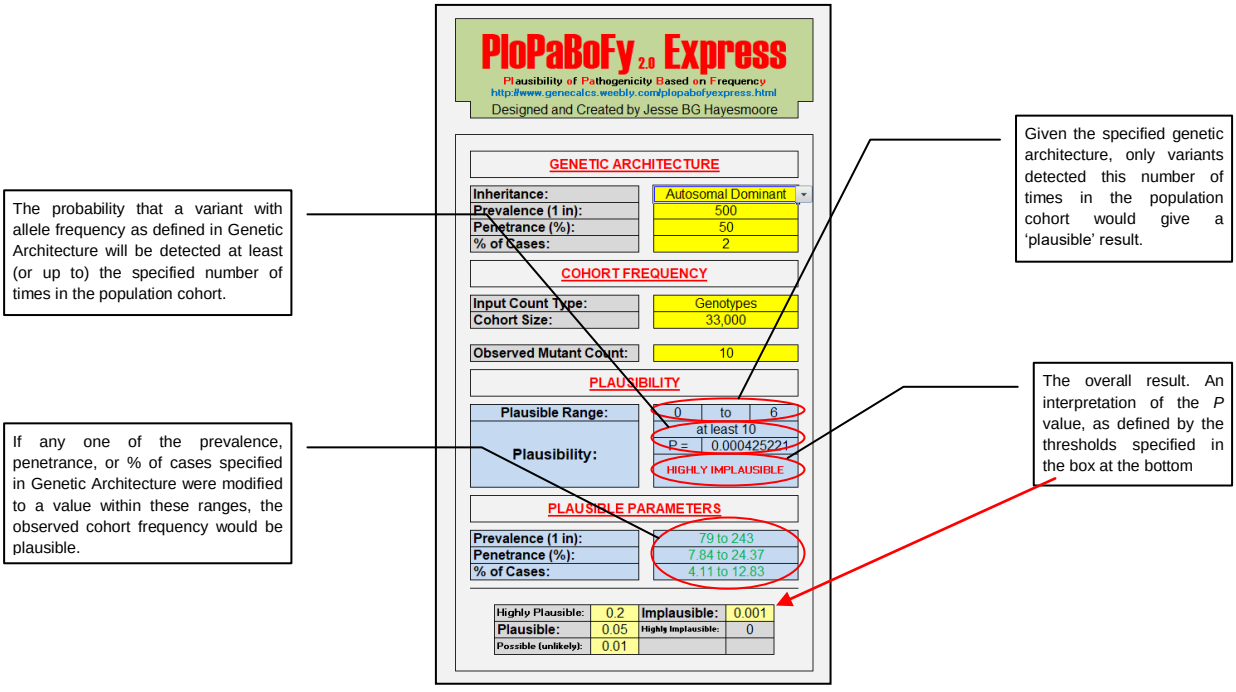
1. **Is variant *A*, detected in *x* of *y* individuals (or alleles), a plausible pathogenic mutation given the expected population frequency for a pathogenic mutation?**
    - In other words, is the population frequency of variant *A* consistent with the specified (i.e. *prior*) genetic architecture?
    - This approach can be used to assess the plausibility that a specific identified genetic variant could account for a proportion of disease cases (or a proportion pathogenic mutations for autosomal recessive disease).

2. **If variant *A*, detected in *x* of *y* individuals, is assumed to be a pathogenic mutation, then what would be a plausible genetic architecture?**
    - In other words, what (*posterior*) genetic architecture would be consistent with the population frequency of variant *A*?
    - In this approach, users can vary any aspect of the genetic architecture (e.g. *prevalence*, *% penetrance*, or *% of cases / mutations*) to see what values would have to be true for the genotype count of variant *A* to be considered a plausible pathogenic mutation.
    - Plausible values for each aspect of the genetic architecture are suggested as part of the outputted results.

Example: Hypertrophic Cardiomyopathy

Hypertrophic cardiomyopathy (HCM) is an autosomal dominant disease. It is estimated to affect 1 in 500 adults. It is genetically heterogeneous with mutations in >20 genes known to be causative. The most common known causative mutation is *MYBPC3* p.(Arg502Trp), which is detected in 1.7-2% of affected individuals. Incomplete penetrance is generally the norm (50% penetrance is a reasonable estimate for the average).

A *MYBPC3* variant is detected in a proband. A check of ExAC shows it has been detected in 10 of 33,000 non-Finnish Europeans. This information is inputted into PloPaBoFy as follows (see yellow boxes):



The interpretation of the output is as annotated above. In this example, to answer **Question 1:** a variant that accounts for 2% of HCM cases with 50% penetrance (as defined in the Genetic Architecture section) could plausibly be detected in 0 to 6 individuals in a population cohort of 33,000 individuals. The chance that it would be detected in at least 10 individuals is very low (i.e. *P* = 0.000425221); hence, in this example, it is highly implausible that a variant detected in 10 of 33,000 individuals could be pathogenic and account for 2% of HCM cases with 50% penetrance.

On the other hand, to answer **Question 2:** let's now assume that the variant detected in 10 of 33,000 individuals is a known pathogenic mutation. Under this assumption (and assuming the inputted prevalence figure is correct), the PloPaBoFy output indicates that this variant could plausibly account for 2% of HCM cases only if its penetrance is 7.84 to 24.37%. Alternatively, if we are confident that this variant is pathogenic and has 50% penetrance, then the output indicates that this can only be true if the variant accounts for a higher percentage of cases (i.e. 4.11 to 12.83%).

Some Important Notes on Usage

- The above example uses PloPaBoFy Express. The use of the non-Express version is very much the same.
- The use of PloPaBofy is very similar regardless of the inheritance pattern selected, but the following differences should be noted:
    - If the inheritance is X-linked, the values inputted into the Genetic Architecture section should be for males only (i.e. prevalence in males, penetrance in males, and percent of male cases accounted for by a single mutation). However, values inputted into the Cohort Frequency section should be for males and females. The number of females in the cohort should also be inputted (even if 0).
    - If the inheritance is Y-linked, PloPaBoFy assumes that the population cohort is made up entirely of males (i.e. the number inputted into 'cohort size' should be for males only).
    - **The most important difference is when the inheritance is autosomal recessive (see later in bold).**
- Note that *P* values displayed as 0 are usually not precisely zero, but very close to zero (e.g. <0.00000000000001).

For genetically heterogeneous diseases / phenotypes:

- In the Genetic Architecture section, all values entered should at least be true to the inheritance pattern. For example, the prevalence of hypertrophic cardiomyopathy (caused by mutations in >20 genes) is 1 in 500, and the vast majority of cases are autosomal dominant. If, however, X-linked HCM was being considered, the prevalence entered should apply to X-linked HCM (e.g. perhaps 1 in 5,000 if 1% of HCM cases are X-linked).
- Ideally, values entered in the Genetic Architecture section should also be true to the gene in the same way. However, since diseases / phenotypes with autosomal dominant, X-linked, Y-linked, or mitochondrial inheritance are usually caused by a single mutation, it is generally acceptable to 'pretend' that all cases are caused by a single gene (e.g. 'all the disease genes are just isoforms of one gene') and that a single mutation (in any gene) accounts for a set number of cases.
- **The crucial difference for autosomal recessive diseases / phenotypes:** is that a single mutation does not account for a set number of cases, because – by definition – these diseases / phenotypes are caused by the combination of two *trans* mutations in the same gene.
    - **Therefore, values entered in the Genetic Architecture section for autosomal recessive diseases / phenotypes should always be true to the inheritance pattern and the gene.**
    - Instead of inputting values for the percent of *cases*, the inputted values should be for the percent of *mutations* accounted for by a single variant.